Deep Learning

Assignment 5:
Spoken Digits Recognition

Instructor: Dr. Mehrandezh
Student: Marzieh Zamani

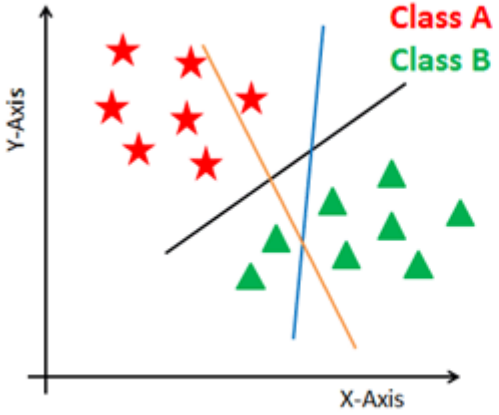
Spoken Digit Recognition on FSDD Dataset

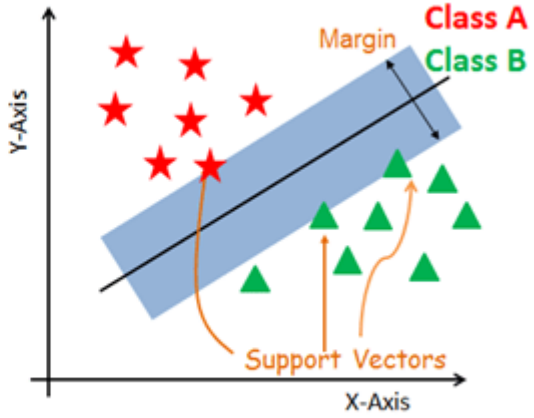
- FSDD Dataset (2 native and 2 non-native speakers)
  - Training: 160 recordings per digit
  - Testing: 40 recordings per digit
- Student Dataset (1 non-native speaker)
  - Evaluation: 5 recordings per digit
- Sampling Frequency: 8000
- Mono

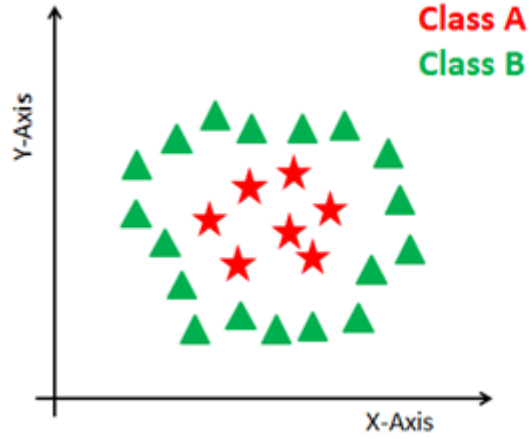
Classifiers

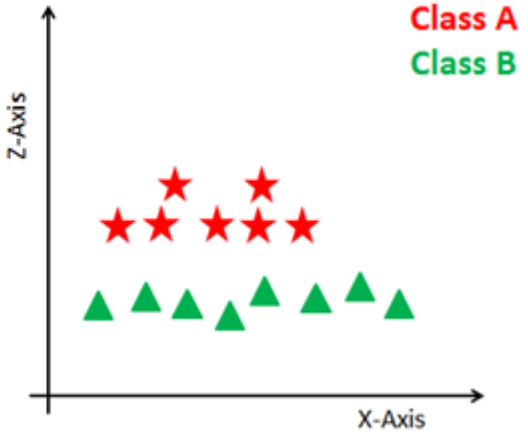
- Support Vector Machine (SVM)
- Long Short-Term Memory (LSTM)
- LSTM optimized by Bayesian Optimization
- Deep Convolutional Neural Network (DCNN) with Mel-frequency Spectrograms

Support Vector Machines (SVM) Classifier





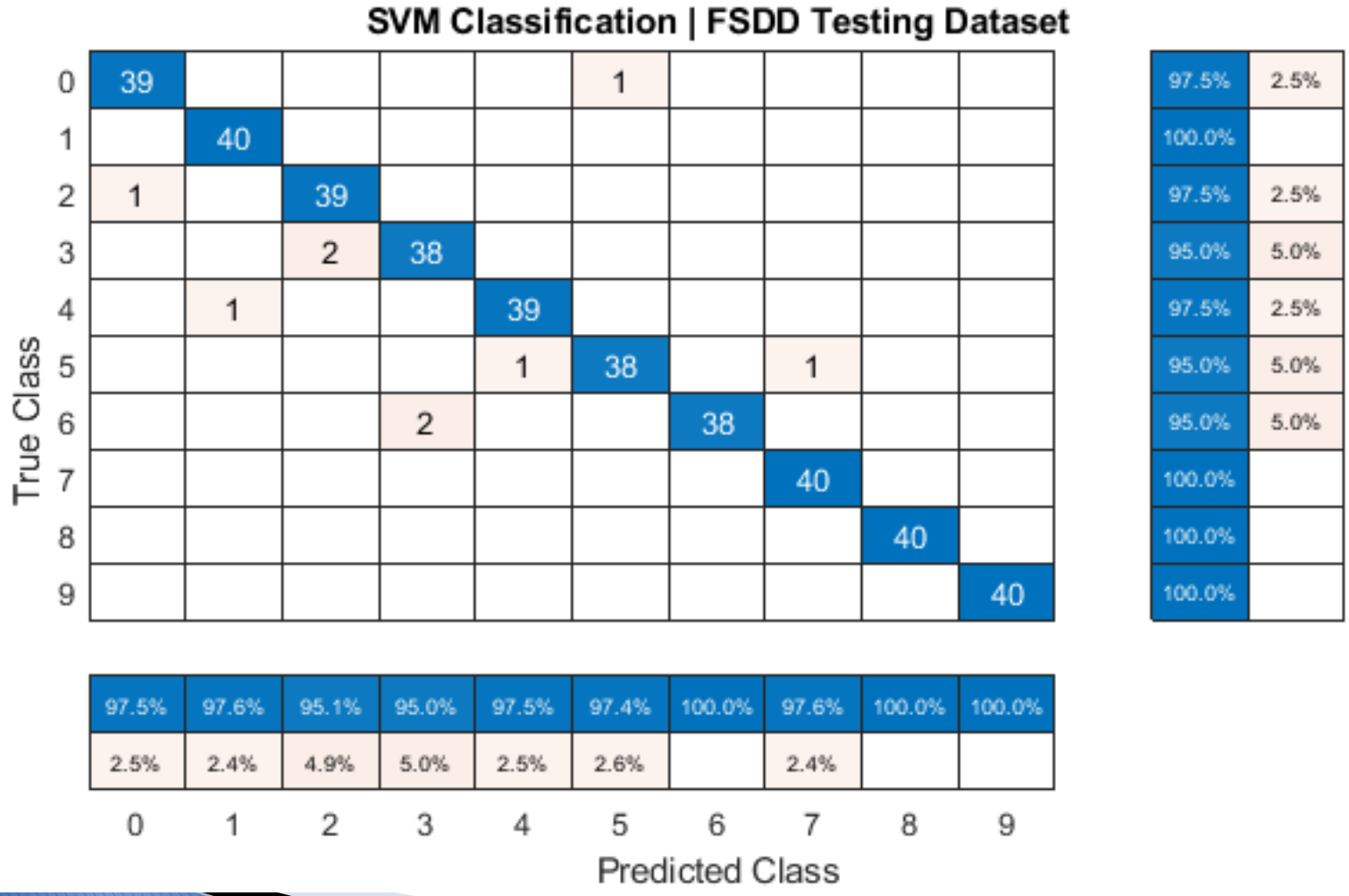




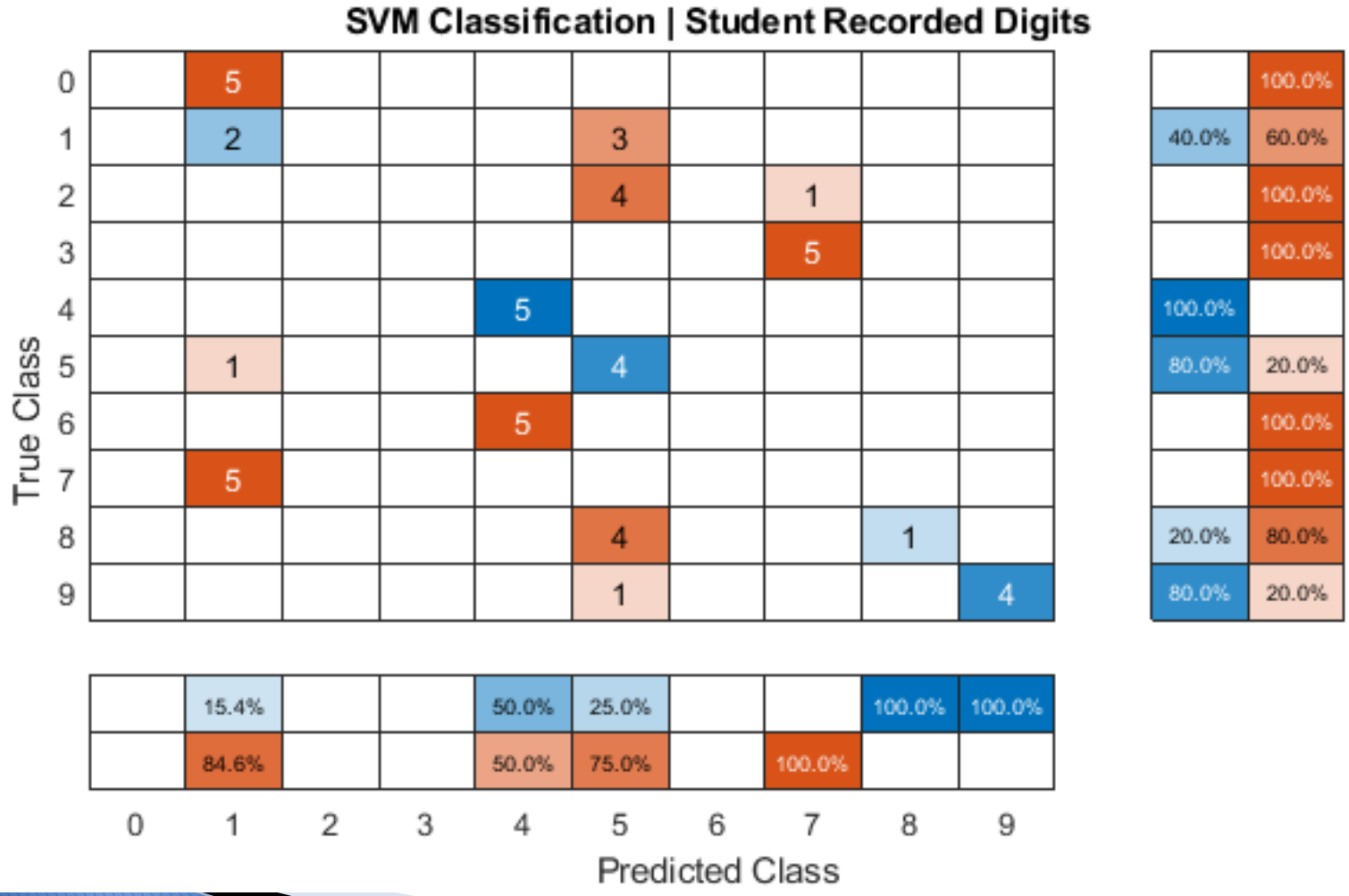
SVM Classifier

- **Specification**: quadratic polynomial kernel
- **Feature vector**: N x 321 x 1
- **Training dataset**: FSDD (N=160 x 10)
- **Testing dataset**: FSDD (N=40 x 10)
- **Evaluation dataset**: Student (N=5 x 10)

SVM Training & Testing on FSDD Dataset => 98%



SVM Evaluation on Student Dataset => 32%

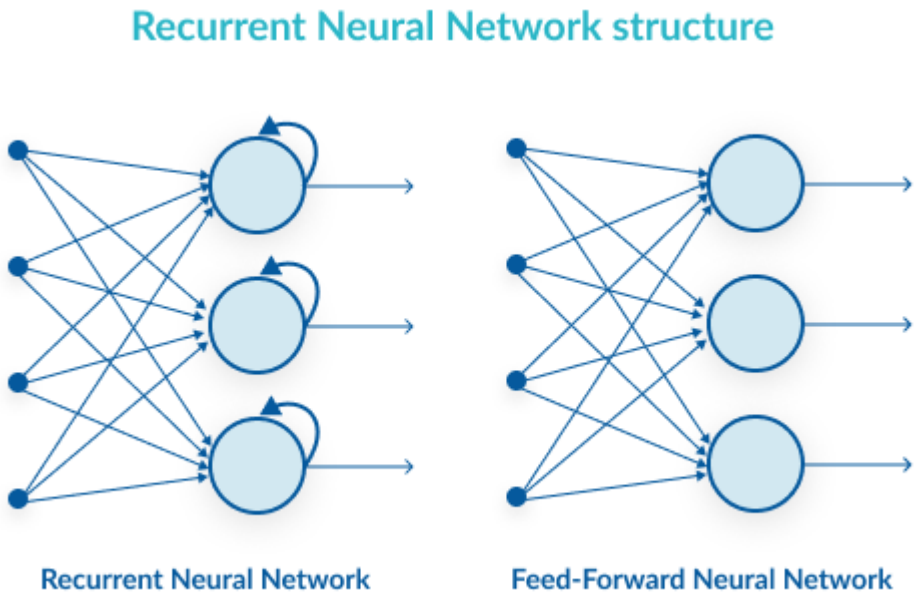


SVM Remarks

- **Testing dataset**: FSDD (N=40 x 10) => test accuracy= 98%
- **Evaluation dataset**: Student (N=5 x 10) => test accuracy= 32%
- SVM classifies FSDD dataset with an excellent accuracy of 98%; however, its classification accuracy for student spoken digits is significantly lower, 32%.
- It is worth mentioning that the classification accuracy of different digits are very distinct, ranging from 0% to 100%.

Feed-forward vs. Recurrent NN

- Feed-forward neural networks pass the data forward from input to output
- RNN has a feedback loop where data can be fed back into the input

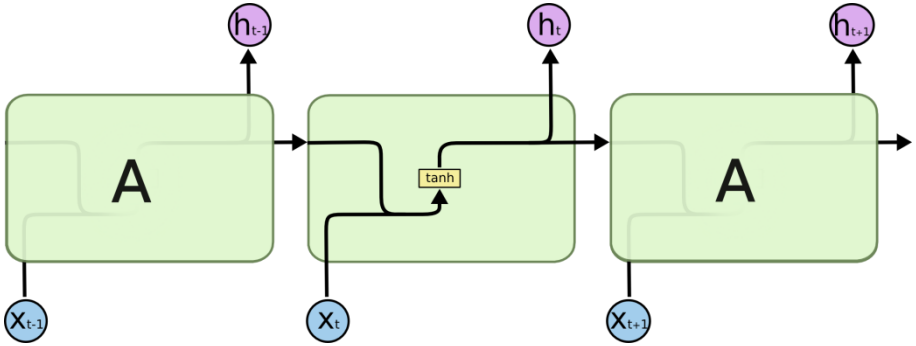


Long Short-Term Memory (LSTM) Classifier

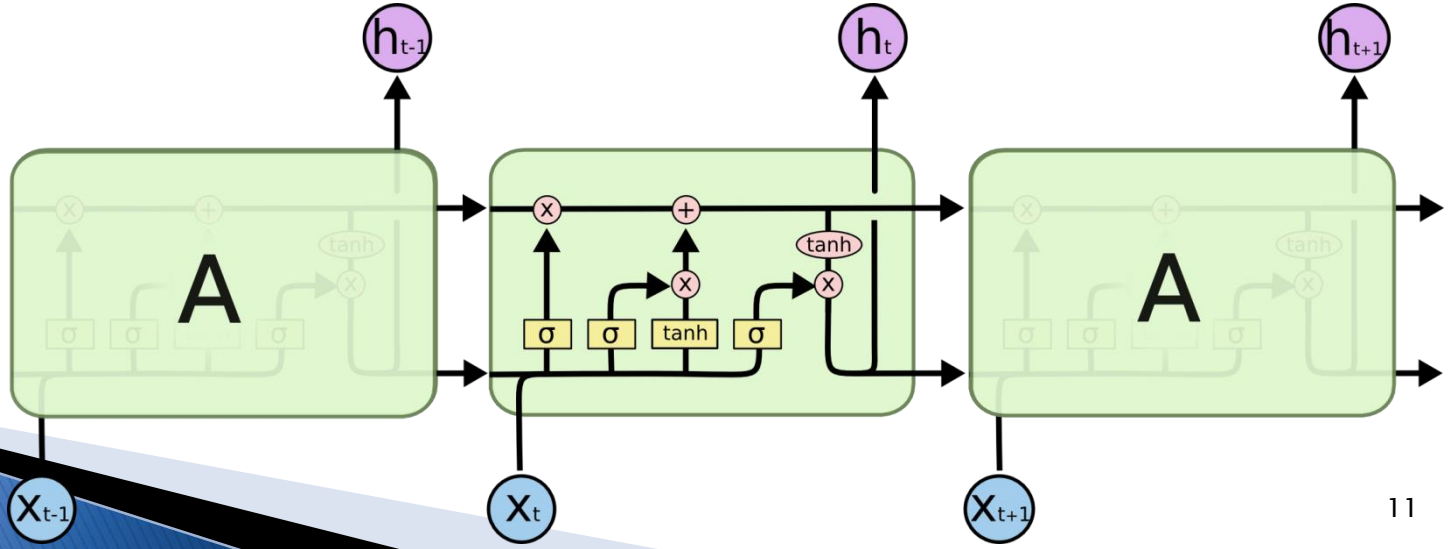
- Designed to remembering information for long periods of time;
- The cell state is kind of like a conveyor belt of information
- Gates: optionally let information to cell state.

Long Short-Term Memory (LSTM) Classifier

- The repeating module in a standard RNN contains a single layer.



- The repeating module in an LSTM contains four interacting layers.

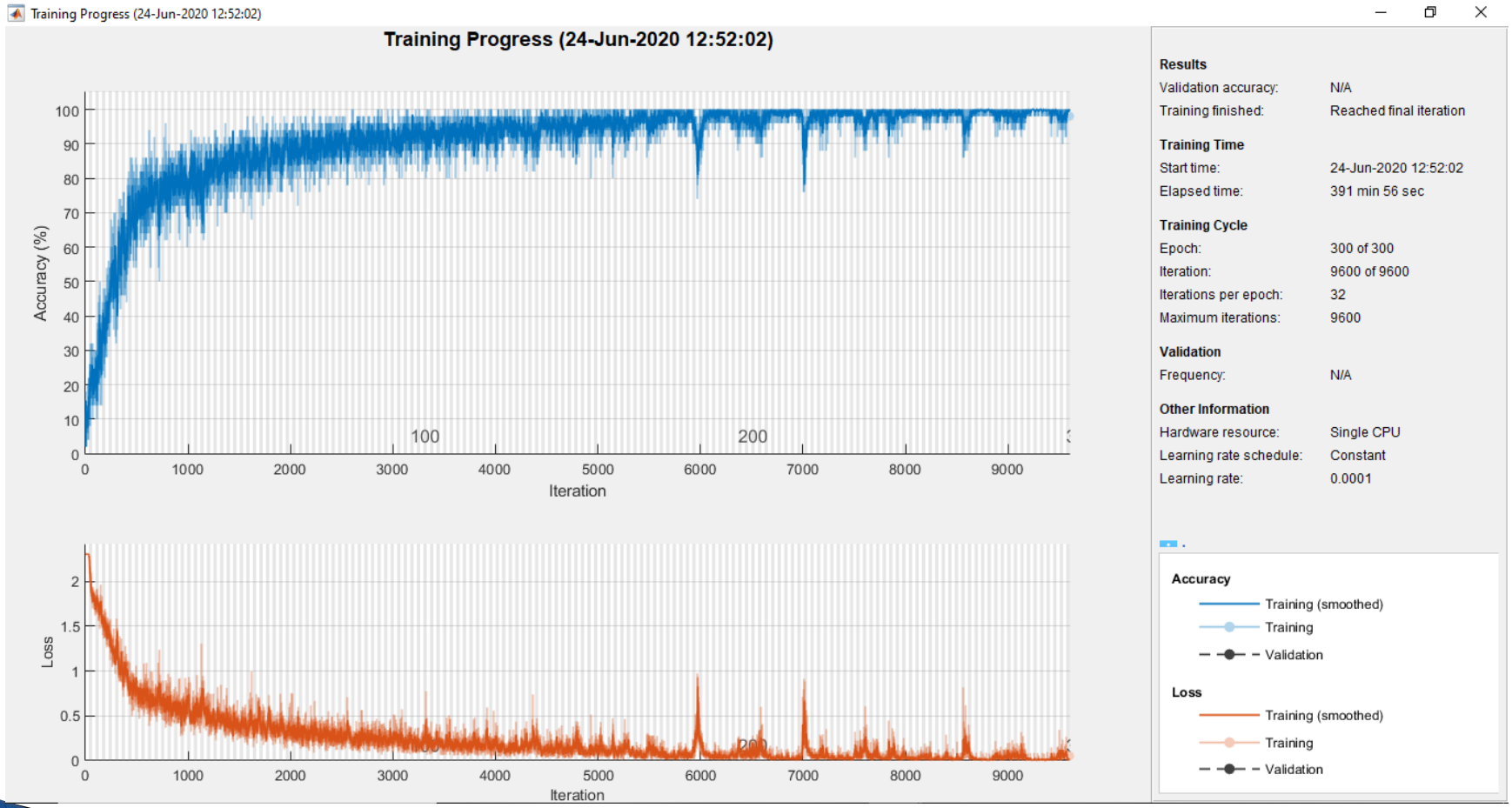


LSTM Architecture & Parameters

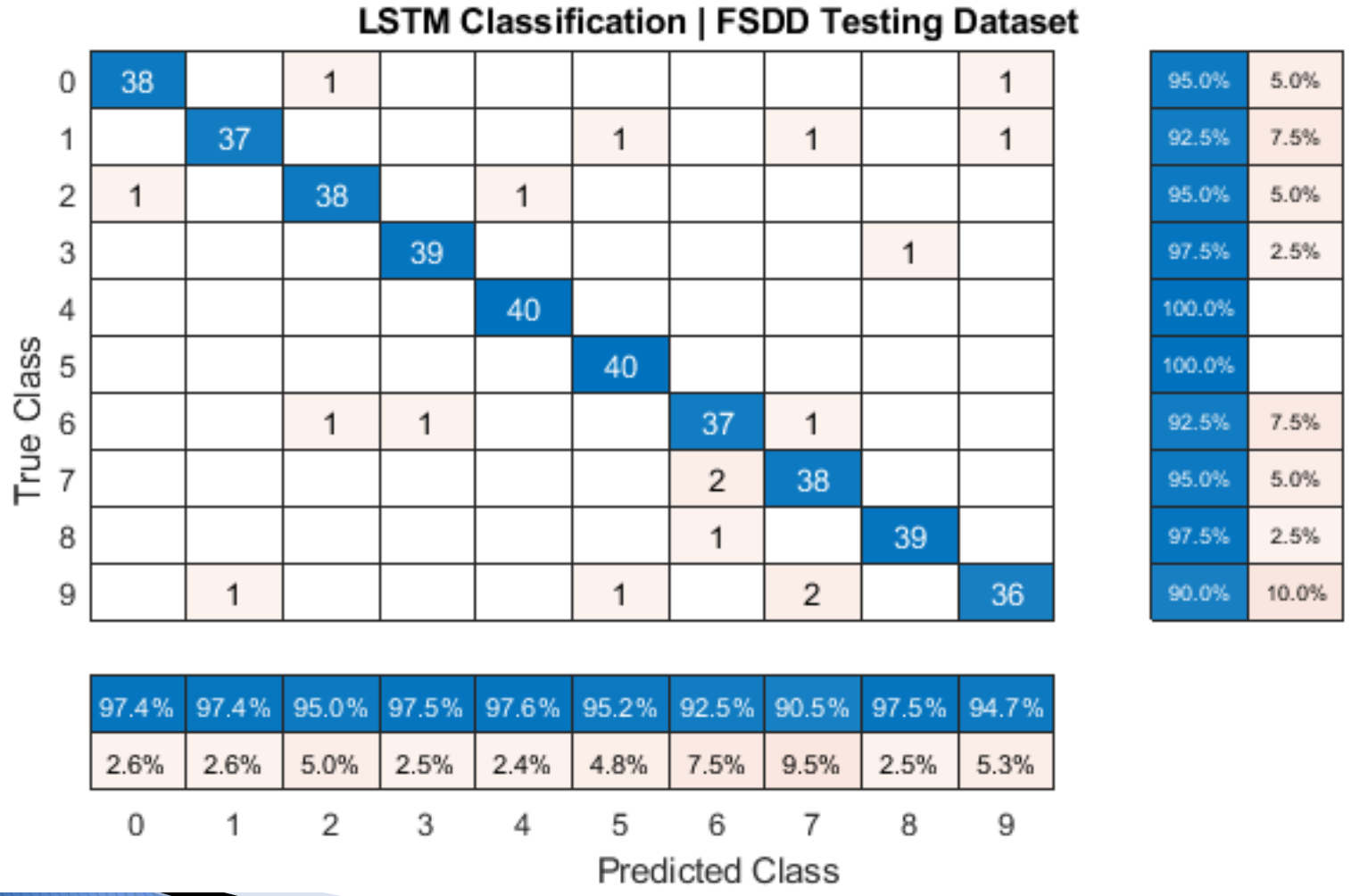
- 512 hidden units
- Initial Learning Rate: 0.0001

No	Layer Type	Output Size	Details
1.	INPUT	321	-
2.	LSTM	512	Input Weights: 2048x321 Recurrent Weights: 2048x512 Bias: 2048x1
3.	FC	10	Weights: 10x512 Bias: 10x1
4.	SOFTMAX	10	-
5.	Class Output	1	-

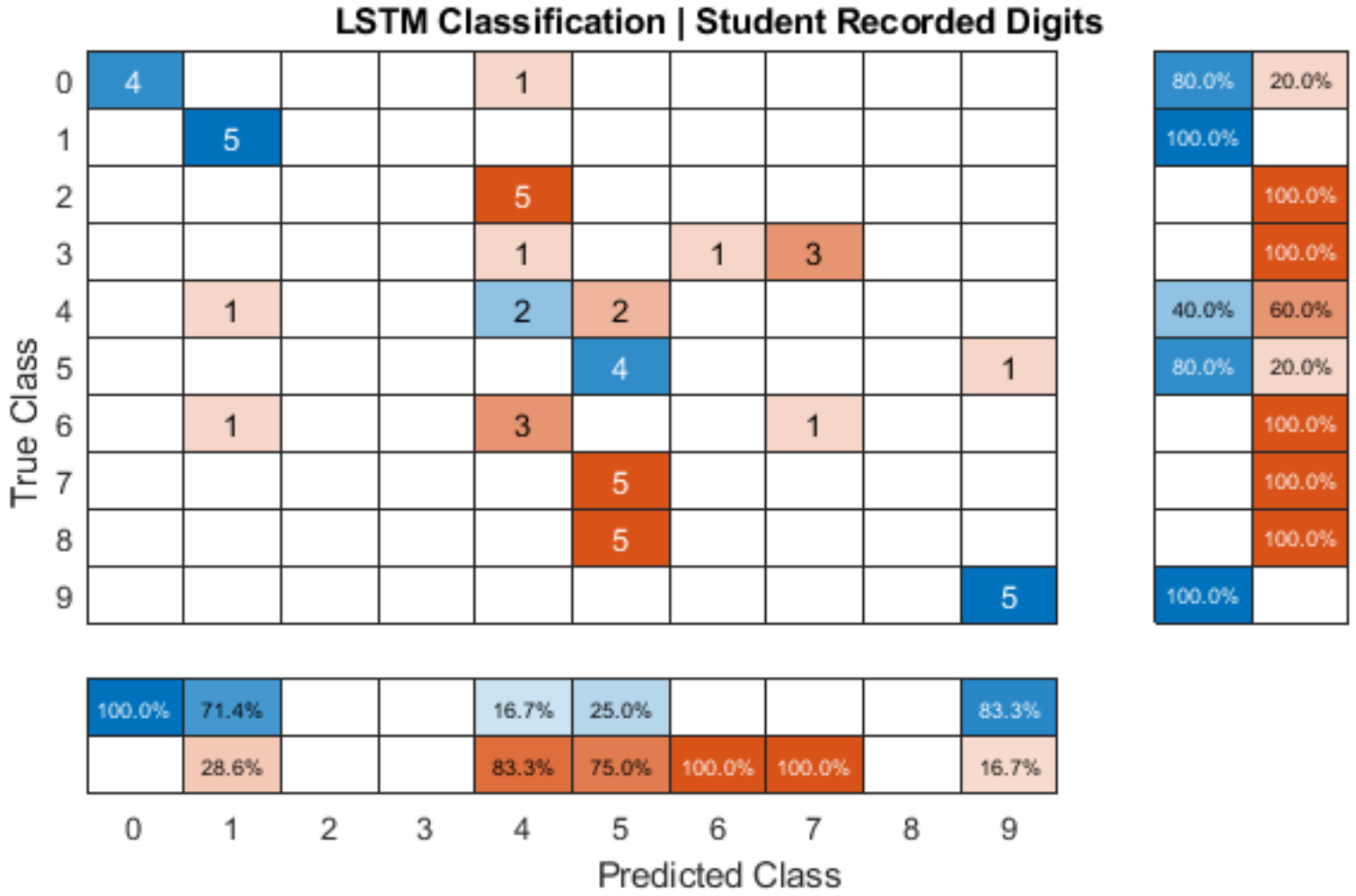
LSTM Training on FSDD Dataset



LSTM Testing on FSDD Dataset => 94%



LSTM Evaluation on Student Spoken Digits => 40%



LSTM Remarks

- **Testing dataset**: FSDD (N=40 x 10) => test accuracy= 94%
- **Evaluation dataset**: Student (N=5 x 10) => test accuracy= 40%
- LSTM classifies FSDD dataset with an excellent accuracy of 94%; however, its classification accuracy for student spoken digits is significantly lower, 40%.
- It is worth mentioning that the classification accuracy of different digits are very distinct, ranging from 0% to 100%.

LSTM-Optimized by Bayesian Optimization

- **Optimization parameters and range**
  - **Initial Learning Rate**: [1e-5, 1e-1]
  => Optimal value = 2.199e-04

  - **Number of Hidden Units:** [10, 1000]
  => Optimal value = 768

- **Feature vector**: N x 321 x 1

- **Training dataset**: FSDD (N=160 x 10)

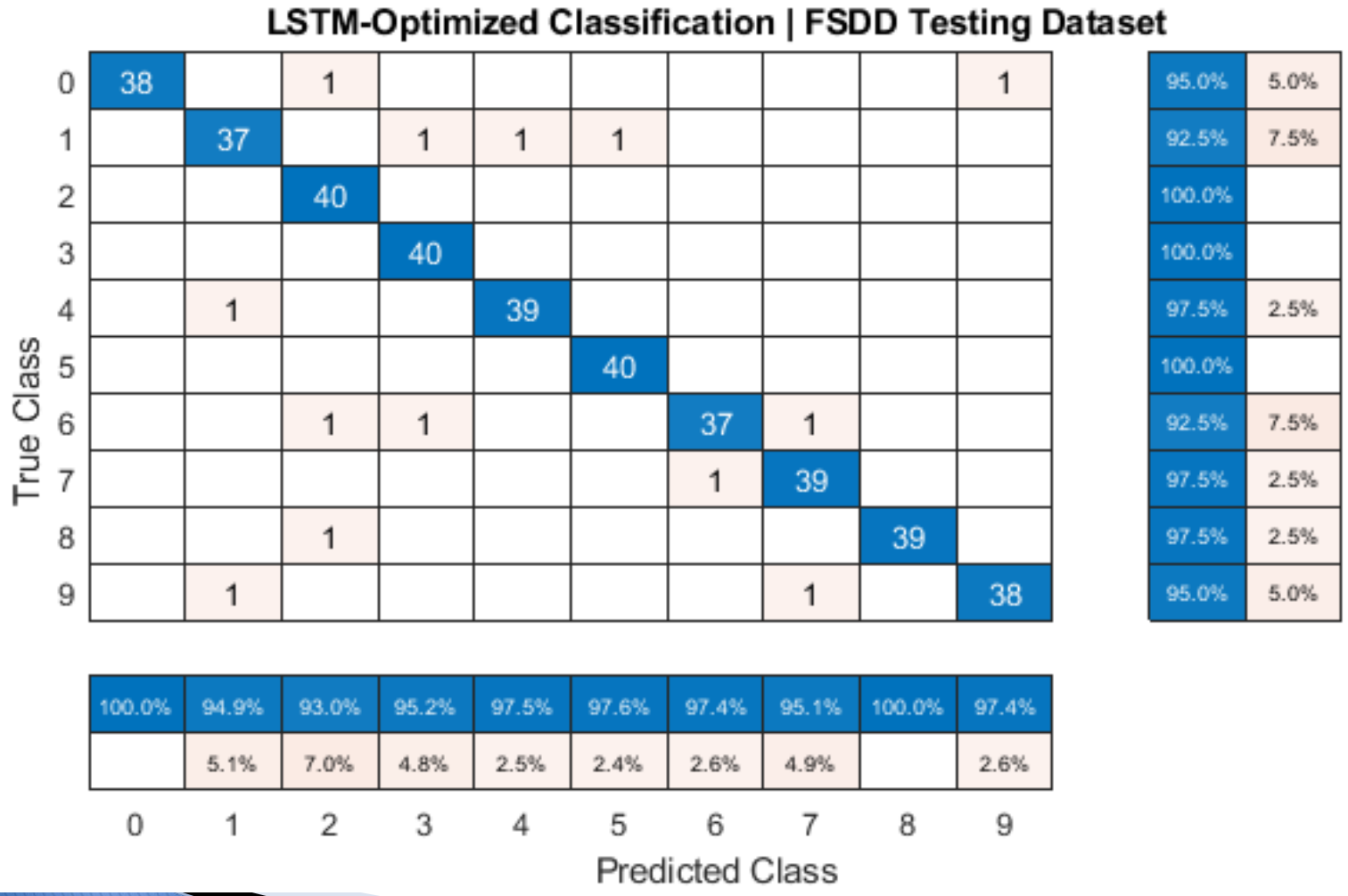
- **Testing accuracy**: FSDD (N=40 x 10)

=> test accuracy: 94% increased to 97.75%

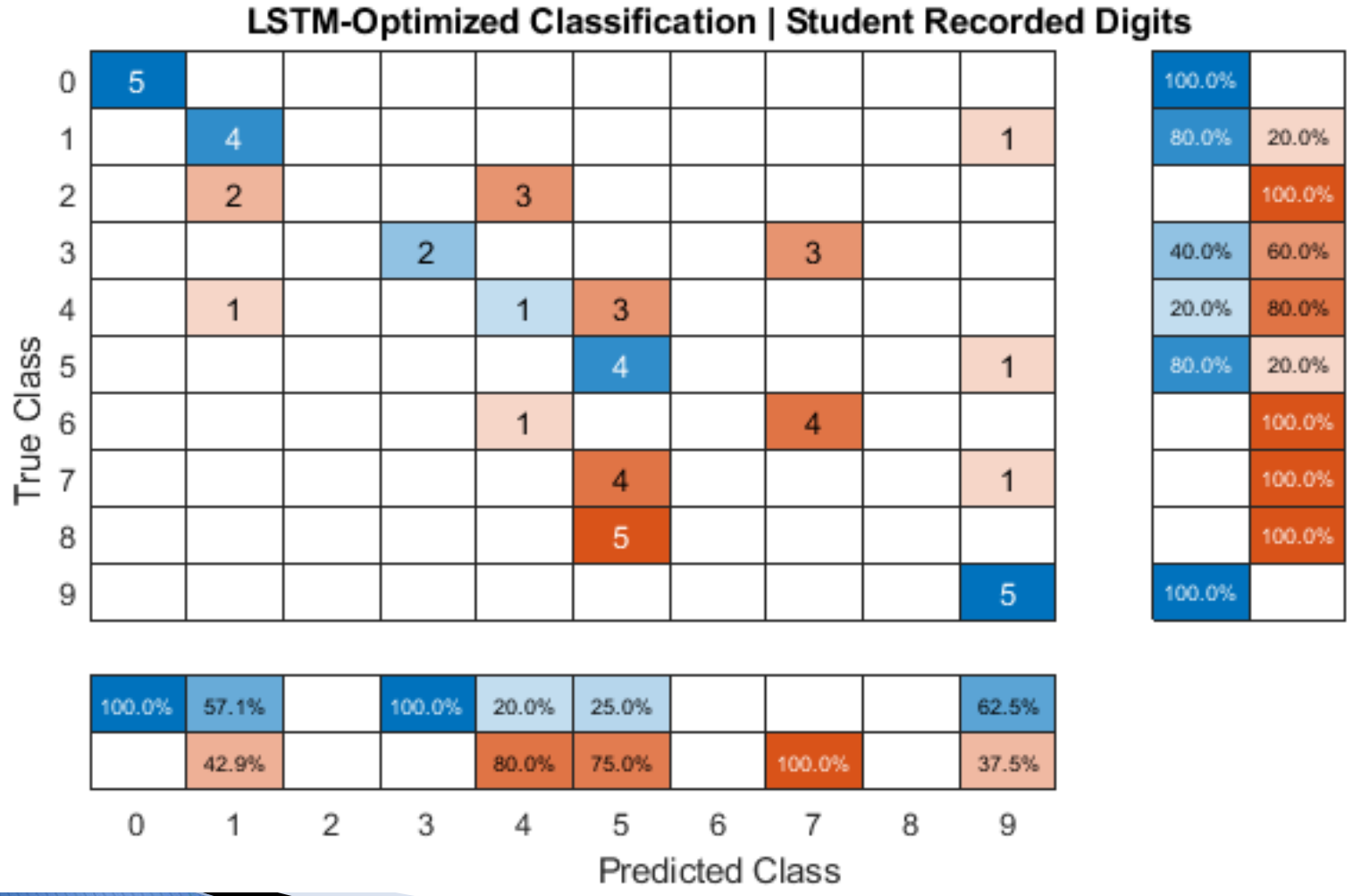
- **Evaluation dataset**: Student (N=5 x 10)

=> test accuracy: 40% increased to 42%

LSTM-Optimized Testing on FSDD Dataset => 98%



LSTM-Optimized Evaluation on Student Spoken Digits => 42%



LSTM-Optimized Remarks

- **Optimization parameters and range**
  - **Initial Learning Rate**: [1e-5, 1e-1]
  => Optimal value = 2.1988

  - **Number of Hidden Units:** [10, 1000]
  => Optimal value = 768

- **Testing accuracy**: FSDD (N=40 x 10)

=> test accuracy: 94% increased to 97.75%

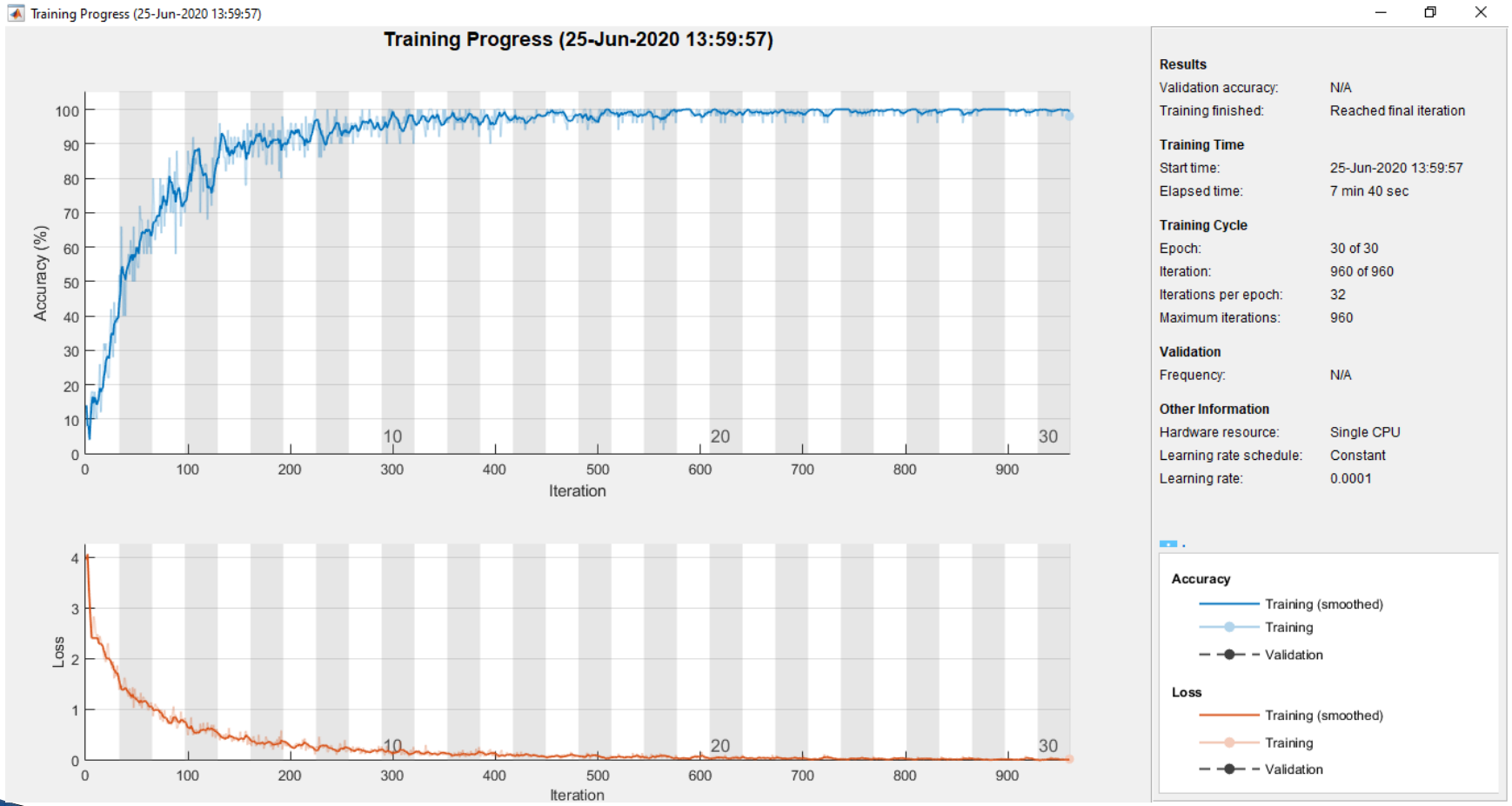
- **Evaluation dataset**: Student (N=5 x 10)

=> test accuracy: 40% increased to 42%

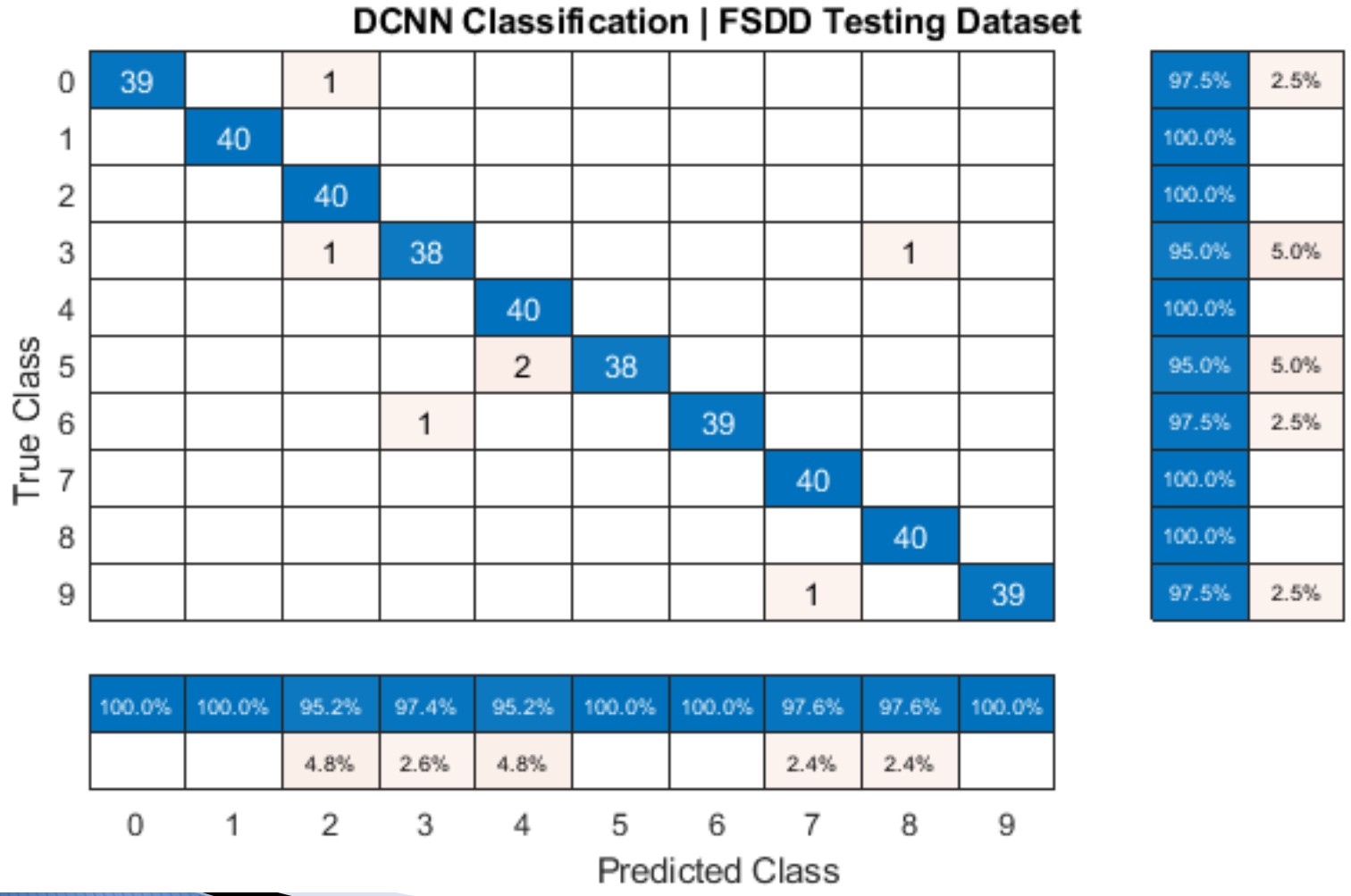
DCNN Architecteur and Parameters

No	Layer Type	Output Size	Filter Size / Stride
1.	INPUT IMAGE	40×81×1	
2.	CONV	40×81×12	5×5;*K* = 12
3.	BN	40×81×12	
4.	ACT (ReLU)	40×81×12	
5.	POOL	20×41×12	F=3, S=2
6.	CONV	20×41×24	3x3;*K* = 24
7.	BN	20×41×24	
8.	ACT (ReLU)	20×41×24	
9.	POOL	10×21×24	F=3, S=2
10.	CONV	10×21×48	3x3;*K* = 48
11.	BN	10×21×48	
12.	ACT (ReLU)	10×21×48	
13.	POOL	5×11×48	F=3, S=2
14.	CONV	5×11×48	3x3;*K* = 48
15.	BN	5×11×48	
16.	ACT (ReLU)	5×11×48	
17.	CONV	5×11×48	3x3;*K* = 48
18.	BN	5×11×48	
19.	ACT (ReLU)	5×11×48	
20.	POOL	4×10×48	F=2, S=1
21.	DO	4×10×48	dropoutProb = 0.2
22.	FC	1x1x10	
23.	SOFTMAX	1x1x10	
24.	Class Output	1x1x1	

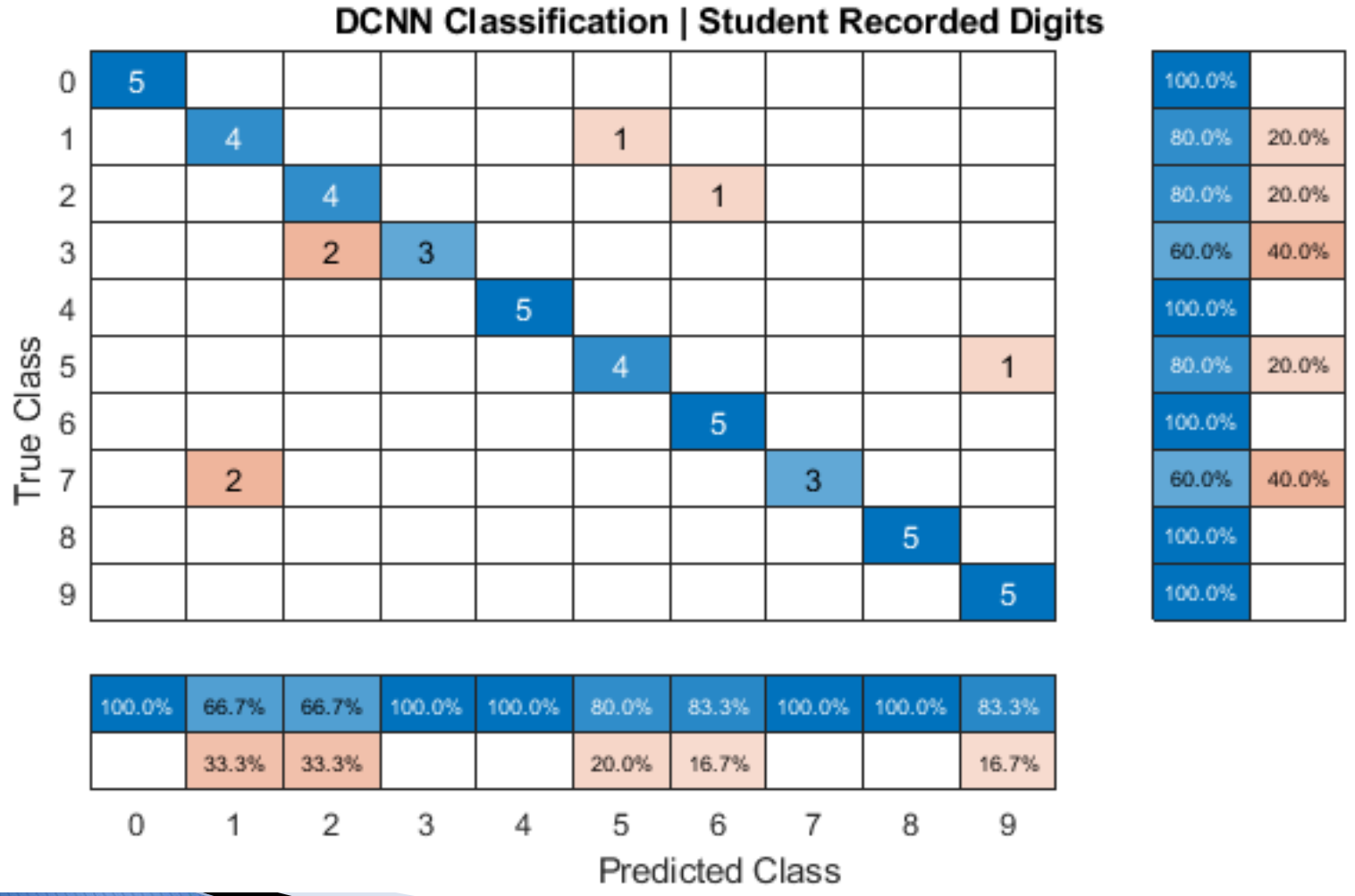
DCNN Training on FSDD Dataset



DCNN Testing on FSDD Dataset => 98%



DCNN Evaluation on Student Spoken Digits => 86%



DCNN Evaluation on Student Spoken Digits

- **Testing dataset**: FSDD (N=40 x 10) => test accuracy= 98%
- **Evaluation dataset**: Student (N=5 x 10) => test accuracy= 86%
- DCNN classifies FSDD dataset with a slightly higher accuracy of 98%.
- Its strength is revealed when it can improve the classification accuracy for student spoken digits from 40% to 86%. This means that not only DCNN is stronger than other networks.

References

1. https://www.mathworks.com/help/audio/examples/spoken-digit-recognition-with-wavelet-scattering-and-deep-learning.html
2. Free Spoken Digit Dataset (FSDD), available at https://github.com/Jakobovski/free-spoken-digit-dataset
3. https://www.mathworks.com/videos/series/understanding-wavelets-121287.html
4. https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python
5. https://colah.github.io/posts/2015-08-Understanding-LSTMs/
6. https://en.wikipedia.org/wiki/Bayesian_optimization#:~:text=Bayesian%20optimization%20is%20a%20sequential,expensive%2Dto%2Devaluate%20functions.

Thanks for your attention

Questions are welcome